Functional Information, Biomolecular Messages and Complexity of BioSequences and Structures (Extended Abstract)

Davide Corona[1], Valeria Di Benedetto[2],
Alessandra Gabriele[2], Raffaele Giancarlo[2],Filippo Utro[2]

[1] Dulbecco Telethon Institute c/o Università di Palermo, Dipartimento di Biologia Cellulare e dello Sviluppo, Palermo, Italy
dcorona@unipa.it

[2] Università di Palermo, Dipartimento di Matematica e Informatica, Palermo, Italy
valeria.dibenedetto@gmail.com,sandra@math.unipa.it,
raffaele@math.unipa.it, utro@math.unipa.it

Abstract. In the quest for a mathematical measure able to capture and shed light on the dual notions of information and complexity in biosequences, Hazen et al. have introduced the notion of Functional Information (FI for short). It is also the result of earlier considerations and findings by Szostak and Carothers et al. Based on the experiments by Charoters et al., regarding FI in RNA binding activities, we decided to study the relation existing between FI and classic measures of complexity applied on protein-DNA interactions on a genome-wide scale. Using classic complexity measures, i.e, Shannon entropy and Kolmogorov Complexity as both estimated by data compression, we found that FI applied to protein-DNA interactions is genuinely different from them. Such a fact, together with the non-triviality of the biological function considered, contributes to the establishment of FI as a novel and useful measure of biocomplexity. Remarkably, we also found a relationship, on a genome-wide scale, between the redundancy of a genomic region and its ability to interact with a protein. This latter finding justifies even more some principles for the design of motif discovery algorithms. Finally, our experiments bring to light methodological limitations of Linguistic Complexity measures, i.e., a class of measures that is a function of the vocabulary richness of a sequence. Indeed, due to the technology and associated statistical preprocessing procedures used to conduct our studies, i.e., genome-wide ChIP-chip experiments, that class of measures cannot give any statistically significant indication about the relation between complexity and function. A serious limitation due to the widespread use of the technology.

Keywords: Functional Activity, Sequence Complexity, Combinatorics on Words, Protein-DNA interaction.

Dagstuhl Seminar Proceedings 10231
Structure Discovery in Biology: Motifs, Networks & Phylogenies
http://drops.dagstuhl.de/opus/volltexte/2010/2688

1 Background

A mathematical theory able to capture the notion of information embodied in a biological system and to describe its complexity is a long sough-after goal, i.e., [5, 30]. Recently, Robert Hazen *et al.* [19] have stressed again the need for such a theory via two important related questions: (1) "What actually is meant by biological complexity?" and (2) "How might that complexity be quantified?".

In the past, several related theoretic methodologies, all revolving around classic notions of sequence complexity, have been proposed in an attempt to find satisfactory answers to the above questions. We limit ourselves to mention some of them, referring the reader to [3, 5, 18, 19] for an in-depth presentation of the state of the art. Gell-Mann [15, 16] gives a crisp formulation of part of the problem and suggests a mathematical solution. Standish [27] discusses biological complexity in terms of Universal Turing machines. Adami and Cerf [4] propose a solution to their particular formulation of the problem by using the concept of entropy in finite ensembles. Galas *et al.* [14] devise a class of measures to quantify the contextual nature of the information in sets of "objects". That new class is based on a generalization to "objects" of Kolmogorov sequence complexity [22].

All of the above contributions to this area seem to have an inherent limitation: the underlying theories ignore the "meaning or function" of a sequence and they attempt to assess its complexity by establishing the existence of concise encodings for the sequence. More in general, although the literature on the above subject is extensive, it seems to be lacking a focus on the relationship between information and function. Concentrating on biopolymers, Szostak [28] outlines a new measure that, departing from previous proposals, quantifies the complexity of a system in terms of the information it needs to acquire in order to develop the ability to perform a given function. That measure is motivated by an experiment conducted by Carothers *et al.* [8] and it is formalized in [19]. It is referred to as *Functional Information* (`FI` for short) and, for convenience of the reader, we define it and briefly mention next the way it has been derived.

1.1 Functional Information

Let Σ be an *alphabet* and let Σ^* be the corresponding *free monoid*. Fix a finite subset U of Σ^*, where each sequence has finite length, and a measurable function F, where the measure is defined over U. U is referred to as *sequence space* (with respect to function F). Let N_k be the number of sequences in U that has a value of the measure at least k. Then, $\mathtt{FI}(k) = -\log_2 \frac{N_k}{|U|}$.

Carothers *et al.* analyze the distribution of functional RNA aptamers in a random population, providing data on a specific example. They identify 11 distinct classes of GTP-binding RNAs, which are distinguished from each other both by nucleotide sequences and secondary structures. The study shows that GTP binding activity to a target molecule requires more complex structural solutions, which, in turn, decrease in sequence space according to an exponential law. They also point out that there exists a good correlation between the growth

of (an *ad hoc* defined) Shannon entropy of the secondary structure of the selected RNA aptamers and their GTP binding affinity. That is, the higher the binding affinity of the RNA aptamer, the higher the entropy of its secondary structure. Figure 1 provides an example. An analogous result can be obtained by using Kolmogorov complexity instead of entropy [23]. The latter is approximated via grammar-based data compression.

Szostak outlined the "desiderata" of a function able to capture biological complexity as one that would establish a relationship between "information content" and biological activity and that should quantify the amount of information necessary to specify a sequence whose activity exceeds a given threshold. Exemplifying via the experiment by Carothers *et al.*, such a function would establish the number of "functional bits" needed for an RNA aptamer to acquire a given level of a specific function: binding affinity with GTP. Among the goals of the measure outlined by Szostak, there is also its use for the identification of common laws, as well differences, governing the general endeavor of "function acquisition" by an aptamer.

Hazen *et al.* present a mathematical formalization of `FI`, reported at the beginning of this section, and they also identify properties of a generic system that are brought to light when one uses `FI` to measure its complexity. Two symbolic systems, in which quantification of `FI` is actually possible, are discussed: alphanumeric sequences and Avida artificial life genomes. This latter is a virtual world populated by digital organism, i.e., computer programs, that self-replicate, mutate, and adapt by natural selection. Analysis of these two systems in terms of `FI` reveals several characteristics that are important in understanding the behavior of systems composed of many interacting agents. First, letter sequences and Avida genomes both display that highly functional configurations comprise only a small fraction of all possible sequences, exhibiting an *exponential decay property* in sequence space analogous to the one displayed by aptamers in the experiment by Carothers at al.. Second, extensive experimentation with the ability of Avida genomes to acquire function, e.g., the ability to perform a given number of arithmetic/logic operation, shows that several discrete classes of functional configurations exist, a situation that yields to distinctive step features in plots of information versus function. That is, a *stepped behavior* of `FI`. Figure 2 provides an example. It is worth pointing out that such a stepped behavior is absent in Avida genomes that perform statistically random functions. Finally, Hazen *et al.* present an intuitive discussion indicating that the results by Carothers *et al.* imply that RNA binding activity, when described by `FI`, also satisfies both the exponential decay and the stepped property.

Hazen *et al.* stress that `FI` may point to key and unifying factors in the origin and emergence of biocomplexity and hence it could be a good candidate to measure biological complexity. However, to date, the only "biological function" that has been described via `FI` is RNA binding activity. Moreover, the research by Hazen *et al.* poses, very naturally, several open problems. We mention that it is open whether the stepped behavior of `FI` in the functional Avida genomes is a feature of `FI` or an idiosyncrasy of that particular system. Second, and

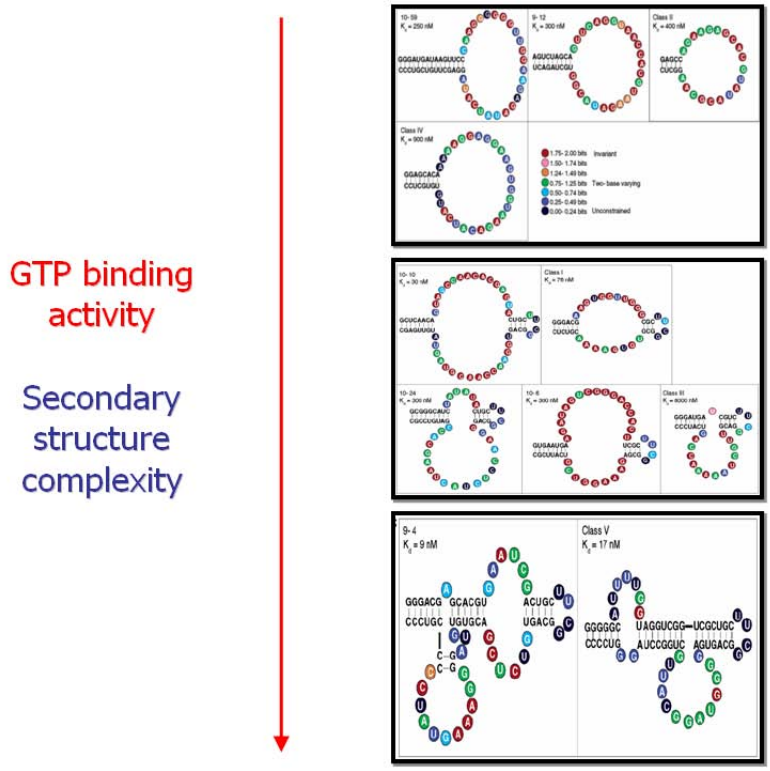

Fig. 1. (Adapted from [8]) A remarkable correspondence between the affinities of the optimized aptamers and the intricacy of their secondary structures. The sequences shown have been optimized for GTP binding. The four simple stem-loop aptamers are relatively weak binders, the five internal bulge-loop aptamers are better binders than the previous one, and the remaining two aptamers are more complex, with three stems and two internal bulge-loops, and they are the best binders. More formally, the figure points out the existence of a good correlation between the growth of (an *ad hoc* defined) Shannon entropy of the secondary structure of the selected RNA aptamers and their GTP binding affinity. The intricacy of the secondary structure of the aptamers can also be formalized by using other complexity measures, e.g. Kolmogorov complexity.

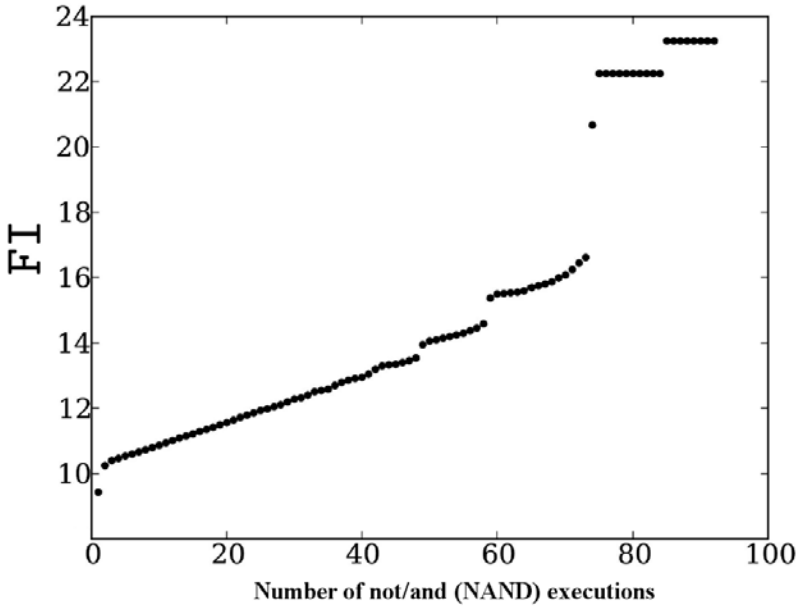

Fig. 2. (Adapted from [19]) The distribution of function of 300-line Avida genomes within a randomly generated sequence space of 10^7 genomes. The abscissa indicates the degree of function E, i.e. the number of times a not/and (NAND) operation is executed by the genome. The ordinate gives the values of `FI` (in bits). Notice that the curve has "steps" in it, indicating that increase in function is not a smooth process.

most important, the definition of FI is strikingly similar to the classic one of self-information [12], a random variable of which entropy is the expectation. In addition, the only experiment involving biological function shows a good degree of correlation between the growth of FI and the growth of information measures related to entropy and Kolmogorov complexity. Although one can easily come up with artificial systems in which such a correlation is lost, it remains open to establish the novelty of FI, with respect to more classic measures of information, via a meaningful biological function.

1.2 Protein-DNA Interaction

We concentrate on protein-DNA interaction, i.e., the biological function consisting of a protein binding in a particular area of a genome, to shed further light on FI. In very broad terms, our experiments show that protein-DNA interaction has the same exponential decay property in sequence space and the same stepped behavior of FI as RNA binding affinity and Avida genomes. Therefore, we add the *second* biological function to the collection of the ones described by FI that exhibits key features of the measure. We also study its relation with classic notions of sequence complexity, namely, entropy and Kolmogorov complexity, as approximated by data compression [17], and our experiments establish that it is genuinely a novel measure of complexity. We also establish an anti-correlation between the growth of the classic complexity measures and the level of function achieved by a genomic region with respect to the binding of proteins. As discussed later, this fact puts on solid ground a heuristic on which motif discovery algorithms are designed. We point out that we have also considered measures that are based on linguistic complexity [6, 21] and classic notions of combinatorics on words [13]. Unfortunately, they do not seem to be well suited for the task because of the nature of the data we consider, rather than the function, as illustrated later.

2 Experimental Methodology

We concentrate on genome-wide studies on protein-DNA interaction, with particular attention to chromatin remodelers in the model organism *Drosophila melanogaster*. In particular, we use 14 ChIP-on-chip experiments, taken from Schuettengruber *et al.* [26] and the modENCODE project [9]. For each experiment, we extract genomic regions of high enrichment via statistical procedures that assign a score to each region. Only scores that provide a small percentage of false positives are considered. Moreover, in order to ensure that our experimental results are not an artefact of the technology or of the statistical procedures employed for the processing of the ChIP-on-chip data, we use the Galaxy peak score [10] for the Schuettengruber dataset and the modENCODE peak score for the remaining datasets [1]. In both cases, we take the value of the peak score given to a genomic region as the measure of protein-DNA interaction. In conclusion, each of our 14 datasets is composed of a set of sequences, where each

sequence has a score assigned to it. It is worth pointing out that, for conciseness, we use only one dataset to highlight our results.

3 Results

3.1 Functional Information

For each dataset, we compute `FI` by setting the sequence space equal to the number of DNA aptamers that have an enrichment factor at least equal to a given minimum score, i.e. nine for the datasets in which we have used the Galaxy peak score. For instance, in Figure 3, the number of sequences exceeding that score is histogrammed. In all of the 14 datasets, we find that such a number exhibits an exponential decay in sequence space. In some cases, the decay is more pronounced than in others. For instance, in Figure 3, it is visible towards the end of the peak score abscissa. Notice that such a behavior of protein-DNA interaction exhibited by our experiments is analogous to the one of RNA binding activity reported by Charoters et al.. Moreover, `FI` exhibits a stepped behavior in most of our experiments. This result is consistent with the analogous ones reported in [19]. Figure 4 provides an example and illustrates the method used to reach that conclusion.

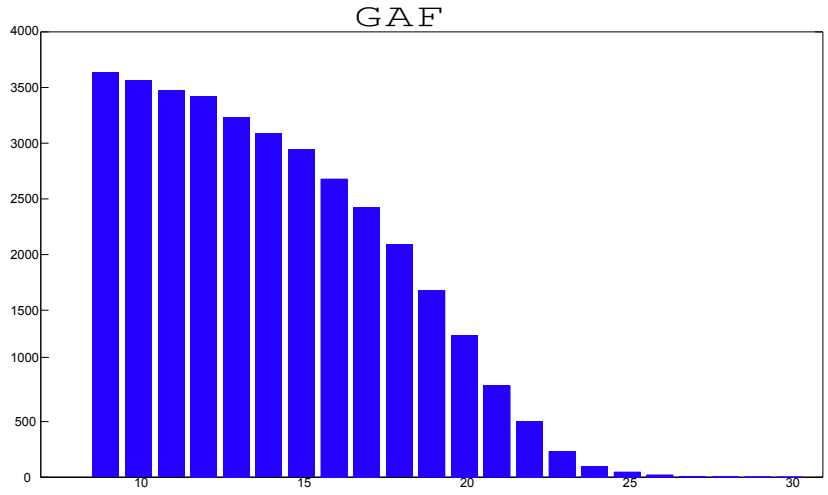

Fig. 3. A histogram of the number of sequences having at least an enrichment factor k, $k \geq 9$, in protein-DNA interaction for the protein GAF, as measured by the Galaxy peak score. The decrease of that number increases with peak score and, towards the end, it has an exponential decay behavior.

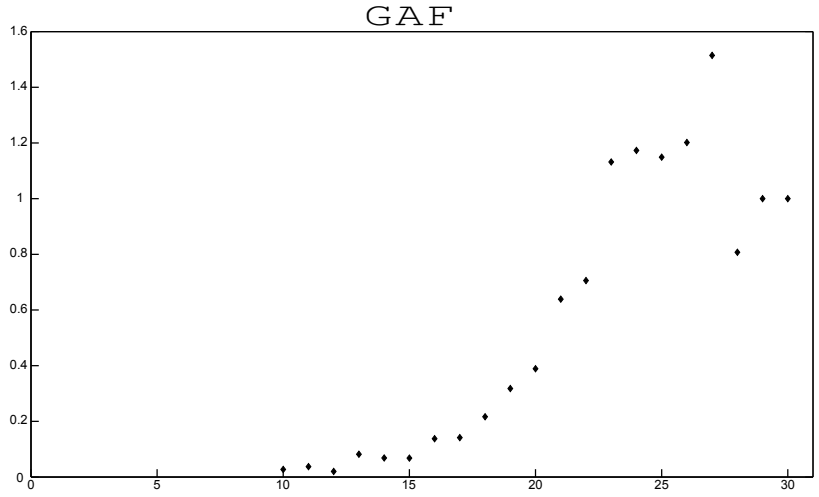

Fig. 4. A stepped behaviour of `FI` for Protein-DNA interaction, again for the GAF protein. In order to better highlight such a behavior, the graph reports, for each integer $i > 9$, the difference between the value of `FI` at i and at $i - 1$. That is, it gives the increment in `FI`, as a function of the measurable activity. In agreement with Figure 3, functional information grows slowly and smoothly at first and then at a more pronounced pace and with two discontinuities.

3.2 Sequence Complexity

We use data compression to compute sequence complexity (SC, for short). As already mentioned, data compression can be seen as an approximation of two related measures of complexity: entropy and Kolmogorov complexity. For our experiments, we use `XM` [7] and `Gencompress` [11], two of the best compression methods available for DNA sequences. We have also used Arithmetic Codes [31] (`AC` for short), which are rather weak general compression routines, since they build simple models of the data. Although the results we have obtained follow the same general trend with all three compressors, the ones obtained with `AC` are the best and we take them as reference. For each sequence, we take the compression ratio as a quantification of its complexity and we refer to this measure as `SC-CR`. In all of our experiments, we find that the "ability" of a genomic area to "interact" with a protein decreases with its lack of redundancy. Figure 5 provides an example. Such a fact is established with the use of Spearman rank correlation test, according to the implementation in the statistical computing environment R [2]. The mentioned result is in agreement with the rule of thumb governing the design of motif discovery algorithms for the identification of binding sites in genomic sequences, e.g., [24], but, prior to our experiments, it had not been assessed and measured on a genomic scale. Indeed, binding sites imply the existence of motifs which, in turn, are an indication of redundancy. The motif discovery algorithms use the following strategy: turn all of the mentioned

implications around and exploit the redundancy in a sequence to identify motifs that *may turn out to be* genuine binding sites. The high degree of correlation between enrichment levels, signaling the presence of true binding events, and the level of redundancy of the corresponding functions state that turning the stated implications around has a very good chance of success on ChIP-on-chip data.

We have also considered combinatorial measures, related to linguistic complexity [20, 29]. In particular, one introduced by De Luca and Varricchio [25] in the realm of formal language theory. It is interesting to report that the ChIP-on-chip sequences extracted as outlined in section 2 are not well suited to be processed by this type of measures. Indeed, the higher the peak score in a genomic area, as established for instance by Galaxy, the longer the sequence. That is, there is an excellent correlation between the peak values and the sequence lengths. Since, in any of our datasets, the sequence lengths are quite different, while measures of linguistic complexity depend on a fixed dictionary, one has that longer sequences get a richer dictionary simply because they have more subsequences in them. Again, there is a high correlation between the richness of the dictionary and sequence length. Therefore, one would expect a high correlation between the linguistic complexity values and the peak scores, as well as `FI`, which is due to the length of the sequences. In fact, in order to avoid this problem, measures of linguistic complexity are used on sequences of equal length in bioinformatics applications [**?**]. The extension of linguistic complexity measures to bioinformatics applications involving sequences of arbitrary length is open.

3.3 `FI` and Sequence Complexity

In our experiments, we find that `FI` is in most cases highly anti-correlated (again with use of Spearman rank test) with `SC-CR`. Figure 6 provides an example. Such an anti-correlation indicates that as the ability of a biosequence to perform a function grows, it is not necessarily true that the complexity of its "combinatorial structure" must grow as well. Such a finding complements quite well the one by Carothers *et al.* and it is a distinctive feature both of the biological function and of the methods that are in use to quantify it. Indeed, the higher the peak score, the higher the density of binding sites of the given protein and, again, the higher the presence of motifs. This latter fact implies redundancy in the genomic region of interest, as already discussed at the end of section 3.2.

4 Conclusions

Our experiments contribute to the establishment of `FI` as a genuinely novel measure of complexity in the realm of biosequences. Moreover, we also highlight a correspondence between the combinatorial richness of a genomic region and its ability to "interact" with a protein. Indeed, motif discovery algorithms are based on the rule of thumb that binding sites imply the presence of motifs which, in turn, imply the presence of redundancy in a biosequence. Our experiments

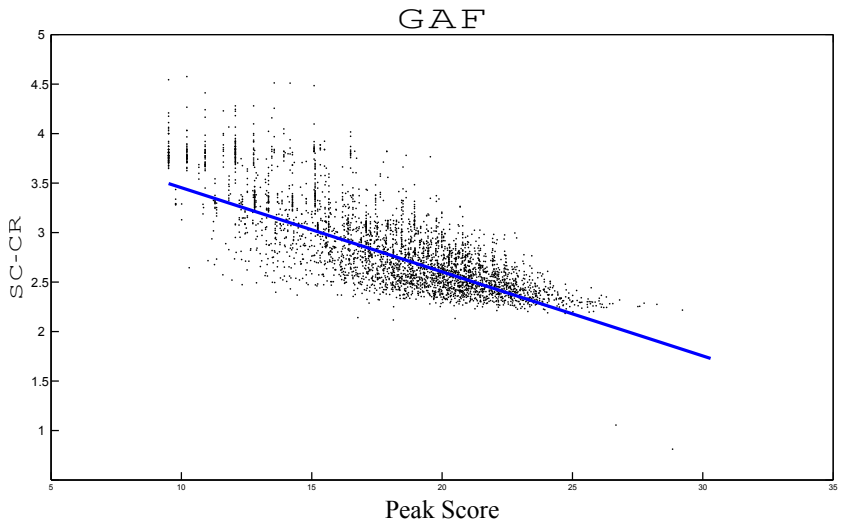

Fig. 5. A plot of the "ability" of a genomic area to "interact" with protein GAF (measured by the Galaxy peak score) and its combinatorial richness (measured by `SC-CR`). The anti-correlation is evident from the plot. The Spearman rank correlation test returns a value of -0.7568949. The Kendall's robust line regression is also shown. Interestingly, the "power of interaction" of a genomic area with GAF decreases with its combinatorial richness, a behavior common to all of the 14 protein-DNA interaction data we have used for our experiments.

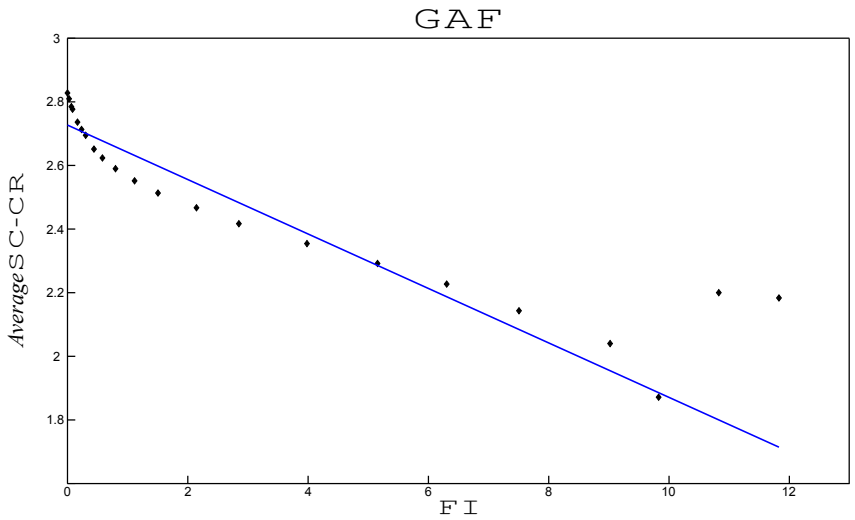

Fig. 6. The relation between `FI` and the *Average* `SC-CR`. Those latter values have been obtained by taking, for each group of sequences with the same value of `FI`, the average over the corresponding values of `SC-CR`. The anti-correlation is evident from the plot. The Spearman rank correlation test returns a value of -0.9830604. The Kendall's robust line regression is also shown. Interestingly, as the level of function grows, the average combinatorial richness of the corresponding sequences decreases, a behavior common to all of the 14 protein-DNA interaction data we have used for our experiments.

provide the first quantitative, genome-wide, positive assessment of that praxis in the design of motif discovery algorithms.

References

1. **The modENCODE Project**. [http://www.modencode.org/].
2. **The R Project for Statistical Computing**. [http://www.r-project.org/].
3. C. Adami. Information Theory in Molecular Biology. *Physics of Life Review*, 1:3–22, 2004.
4. C. Adami and N. J. Cerf. Physical complexity of symbolic sequences. *Physica D*, 137:62–69, 2000.
5. J. C. Avise and F. J. Ayala. In the light of evolution I: Adaptation and complex design. *Proc. of Nat. Acad. Sci*, 104:8563–8566, 2007.
6. A. Bolshoy. DNA sequence analysis linguistic tools: contrast vocabularies, compositional spectra and linguistic complexity. *Appl. Bioinformatics*, 2:103–112, 2003.
7. M. D. Cao, T. I. Dix, L. Allison, and C. Mears. A simple statistical algorithm for biological sequence compression. In *Proc. of the IEEE Data Compression Conference (DCC)*, pages 43–52. IEEE Computer Society, 2007.
8. J. M. Carothers, S. C. Oestreich, J. H. Davis, and J. W. Szostack. Informational complexity and functional activity of RNA structures. *J. AM. CHEM. SOC.*, 126:5130–5137, 2004.
9. S. E. Celniker, L. A. Dillon, M. B. Gerstein, K. C. Gunsalus, S. Henikoff, G. H. Karpen, M. Kellis, E. C. Lai, J. D. Lieb, D. M. MacAlpine, G. Micklem, F. Piano, M. Snyder, L. Stein, K. P. White, and R. H. Waterston. Unlocking the secrets of the genome. *Nature*, 459(7249):927–30, Jun 18 2009.
10. M. Cesaroni, D. Cittaro, A. Brozzi, P.G. Pelicci, and L. Luzi. CARPET: a web-based package for the analysis of chip-chip and expression tiling data. *Bioinformatics*, 24(24):2918–2920, 2008.
11. X. Chen, S. Kwong, and M. Li. A compression algorithm for DNA sequences and its applications in Genome comparison. In *RECOMB 00: Proc. of the 4th Annual International Conference on Computational Molecular Biology*, pages 107–117. ACM, New York, 2000.
12. T. M. Cover and J. A.Thomas. *Elements of Information Theory*. Wiley-Interscience, New York City, 1991.
13. M. Lotaire (Ed.). *Combinatorics on Words*. Cambridge University Press, 1997.
14. D. J. Galas, M. Nykter, G. W. Carter, N. D. Price, and I. Shmulevich. Biological information as set-based complexity. *IEEE Transactions on Information Theory*, 56:667–677, 2010.
15. M. Gell-Mann. *The quark and the jaguar: adventures in the simple and the complex*. New York: W.H. Freeman, 1994. p. 392.
16. M. Gell-Mann and S. Lloyd. Information measures, effective complexity, and total information. *Complexity*, 2:44–52, 1996.
17. R. Giancarlo, D. Scaturro, and F. Utro. Textual data compression in computational biology: a synopsis. *Bioinformatics*, 25(13):1575–1586, 2009.
18. P. Godfrey-Smith and K. Sterelny. Biological information. *The Stanford Encyclopedia of Philosophy*, 2008. Stanford University Press.
19. R. M. Hazen, P. L. Griffin, J. M. Carothers, and J. W. Szostak. Functional information and the emergence of biocomplexity. *Proc. of Nat. Acad. Sci*, 104:8574–8581, 2007.

20. A. K. Konopka. Sequences and codes: fundamentals of biomolecular cryptolofy. *D. Smith, Editor, Biocomputing: Informatics and Genome Projects, Academic Press*, page pp. 119174, 1994. San Diego.
21. A. K. Konopka. Information theories in molecular biology and genomics. *Nature Encyclopedia of the Human Genome*, 3:464–469, 2005.
22. M. Li and P. M. B. Vitányi. *An introduction to Kolmogorov Complexity and its Application*. Springer-Verlag, New York City, 1997.
23. Q. Liu, Y. Yang, C. Chen, J. Bu, Y. Zhang, and X. Ye. RNACompress: Grammar-based compression and informational complexity measurement of RNA secondary structure. *BMC Bioinformatics*, 9:176+, 2008.
24. X. S. Liu, D. L. Brutlag, and J. S. Liu. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol.*, 20:835–839, 2002.
25. A. De Luca and S. Varricchio. *Finiteness and Regularity in Semigroups and Formal Languages*. Springer, 1999.
26. B. Schuettengruber, M. Ganapathi, B. Leblanc, M. Portoso, R. Jaschek, B. Tolhuis, M. van Lohuizen, A. Tanay, and G. Cavalli. Functional anatomy of polycomb and trithorax chromatin landscapes in *Drosophila* embryos. *PloS Biology*, 7(1):e1000013, 2009.
27. R. K. Standish. On complexity and emergence. *Complexity International*, 9:1–5, 2001.
28. J. W. Szostak. Functional information: molecular messages. *Nature*, 423, 2003.
29. O. G. Troyanskaya, O. Arbell, Y. Koren, G. M. Landau, and A. Bolshoy. Sequence complexity profiles of prokaryotic genomic sequences: A fast algorithm for calculating linguistic complexity. *Bioinformatics*, 18(5):679–688, 2002.
30. J. C. Venter and *et al.* The sequence of the human genome. *SCIENCE*, 291, 2001.
31. I. H. Witten, R. M. Neal, and J. G. Cleary. Arithmetic coding for data compression. *Communications of the ACM*, 30:520–540, 1987.